This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

Transportation mode recognition fusing wearable motion, sound and vision sensors

Sebastien Richoz, Lin Wang, Philip Birch, and Daniel Roggen, *Member, IEEE*

***Abstract*—We present the first work that investigates the potential of improving the performance of transportation mode recognition through fusing multimodal data from wearable sensors: motion, sound and vision. We first train three independent deep neural network (DNN) classifiers, which work with the three types of sensors, respectively. We then propose two schemes that fuse the classification results from the three mono-modal classifiers. The first scheme makes an ensemble decision with fixed rules including Sum, Product, Majority Voting, and Borda Count. The second scheme is an adaptive fuser built as another classifier (including Naive Bayes, Decision Tree, Random Forest and Neural Network) that learns enhanced predictions by combining the outputs from the three mono-modal classifiers. We verify the advantage of the proposed method with the state-of-the-art Sussex-Huawei Locomotion and Transportation (SHL) dataset recognizing the eight transportation activities: Still, Walk, Run, Bike, Bus, Car, Train and Subway. We achieve F1 scores of 79.4%, 82.1% and 72.8% with the mono-modal motion, sound and vision classifiers, respectively. The F1 score is remarkably improved to 94.5% and 95.5% by the two data fusion schemes, respectively. The recognition performance can be further improved with a post-processing scheme that exploits the temporal continuity of transportation. When assessing generalization of the model to unseen data, we show that while performance is reduced - as expected - for each individual classifier, the benefits of fusion are retained with performance improved by 15 percentage points. Besides the actual performance increase, this work, most importantly, opens up the possibility for dynamically fusing modalities to achieve distinct power-performance trade-off at run time.**

***Index Terms*—Human activity recognition; transportation mode recognition; data fusion; machine learning; mobile sensing; wearable computing**

I. Introduction

The mode of transportation or locomotion is an important contextual about users during travel, including things such as walking, running, cycling, taking a bus, driving a car, etc [1], [2]. The knowledge of the transportation mode assists context-aware applications, such as activity and health monitoring, individual environmental impact monitoring, and intelligent service adaptation [3]–[9].

Manuscript received: April 8, 2020

S. Richoz, P. Birch and D. Roggen are with the Wearable Technologies Laboratory, University of Sussex, Brighton, UK. L. Wang is with Centre for Intelligent Sensing, Queen Mary University of London, London, UK (e-mail: sr569@sussex.ac.uk, lin.wang@qmul.ac.uk, p.m.birch@sussex.ac.uk, daniel.roggen@ieee.org).

This work was supported by HUAWEI Technologies within the project "Activity Sensing Technologies for Mobile Users". We acknowledge NVidia for GPU donation. L. Wang acknowledges the support from the Institute of Coding, which is supported by the Office for Students (OfS) and the Higher Education Funding Council for Wales (HEFCW).

Nowadays, users carry a growing variety of wearable devices during travel. Besides ubiquitous smartphones, it is ever more common to wear a smartwatch (some of them have integrated cameras), smart earbuds with microphones, bodyworn cameras such as life-loggers or even eye-wear computers (e.g. Google Glass, Spectacles by Snap). These devices are embedded with multimodal sensors including motion sensors, GPS (global positioning system), microphones and cameras. There have been many studies on analyzing the mode of transportation from the data captured by the sensors of these wearable devices with machine learning techniques [10]. Motion and GPS sensors are widely used for transportation mode detection. Motion sensors retrieve the orientation and vibration information of the mobile device while the GPS sensors capture the speed and trajectory of the user [11]–[16]. In comparison to continuous GPS sensing, motion sensors are more desirable as they are much less energy demanding. The state of the art in motion-based transportation recognition performance was established in the SHL recognition challenge 2018 through an open international competition among 20 research teams [17], [18]. The outcomes reveal that approaches based on motion sensors struggle distinguishing between distinct transportation modes of similar kinds: for example between train and subway (rail transport) or between bus and car (road transport).

Sound and vision are two important modalities that are available in wearable devices and can also be used to infer the user's context, although their application to recognizing the mode of transportation has been rarely reported. For instance, a recent challenge on detection and classification of acoustic scenes and events (DCASE) aims to classify various sound events in domestic and wild environments [19], [20]. There has been an increasing number of work using wearable cameras for life-logging, i.e. to recording surrounding environments and the daily life activities of people [21], [22]. The performance of visual object detection and acoustic event classification has progressed significantly since the introduction of deep-learning techniques. In addition, sound, vision and motion are complementary to each other as they each focus on different aspects of user context, providing a high diversity of knowledge.

Many machine learning approaches have been proposed to fuse multimodal information for classification tasks [31], [33], [39], [40]. These approaches can be categorized as early integration (data-layer fusion), late integration (decision-layer fusion). The *early integration* method usually concatenates the data of all modalities as a single input vector for classification, and thus only needs a single classification

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

model. The *late integration* method trains a separate classifier for each modality independently, and draws a final decision by combining outputs of the classifiers. The early integration method considers the cross-modal correlations from the initial stages, and thus potentially outperforms the late integration method, which does not share representations across different modalities and ignores the correlated characteristics among the modalities. However, the synchronization of multiple modalities and the handling of different data size and sampling rate remain an open problem of early integration methods. Furthermore, early integration methods do not easily allow for dynamically changing combinations of sensors: a classifier would need to be trained for each combination of sensors, which limits the scalability of the approach. Late integration methods are inherently modular: each separate classifier is optimized to the corresponding modality, which brings additional benefits of flexibility and scalability.

The motivation behind this paper is twofold. First, we are chiefly interested in investigating how the combination of the three sensor modalities may produce better recognition performance than using a single modality. Second, we are interested in modular approaches, i.e. approaches which enable seamlessly to combine one or more modalities together. Such approaches are important as they enable a system to combine dynamically modalities at runtime, as a way to achieve potentially changing power and performance trade-offs [38]. So far, the combination of motion, sound and vision captured from on-body sensors has not been systematically explored for the recognition of modes of transportations. Due to privacy issues, few transportation and locomotion datasets are publicly available with sound and vision modalities. Only a few work has been reported on transportation mode recognition with vision [23] or sound [24]–[26], and to our knowledge no work has addressed the combination of vision or sound with each other and with motion.

In this paper we conduct the first work that combines the motion, sound and vision modalities for transportation mode recognition. The state-of-the-art Sussex-Huawei Locomotion-Transportation dataset [10], [32] contains rich sensor modalities (including the three above mentioned sensors), which enable us to carry out this research, in order to recognize eight modes of transportation: being still, walking, running, cycling, being in a car, being on a bus, train or subway (Sec. II). Since the three modalities are captured with different sampling rates, the sensor data are not precisely synchronized, and we are interested in modular fusion which can be used in the future for dynamic power/performance management, we focus here on the late integration method. We first train three mono-modal deep neural network (DNN) classifiers, using the motion, sound and vision data, independently (Sec. III). We then evaluate two sets of modular data fusion schemes (ensemble decision and adaptive fusion) that fuse the classification results from mono-modal classifiers (Sec. IV). We compare the performance of combining different modalities with the SHL dataset at the task of recognizing the eight transportation modes (Sec. V). Experimental results demonstrate clear advantages of multimodal fusion. We further assess the generality of the proposed method to unseen data, particularly data comprising user variations (Sec. VI). After discussion in Sec. VII, we draw conclusions in Sec. VIII.

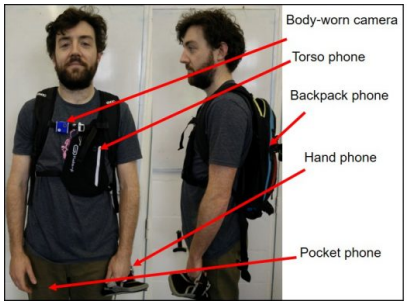


Fig. 1: The equipment for SHL data collection has 4 smartphones and 1 camera. We use the data collected from the hand phone and the body-worn camera.

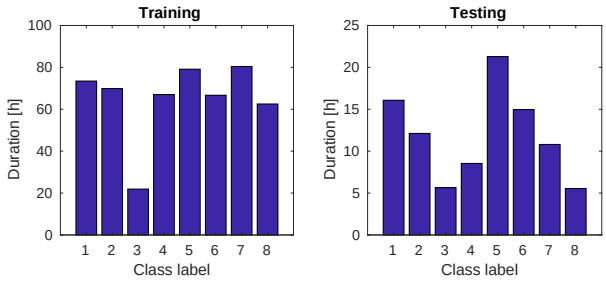


Fig. 2: The duration of each class activity in the training and the testing dataset. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

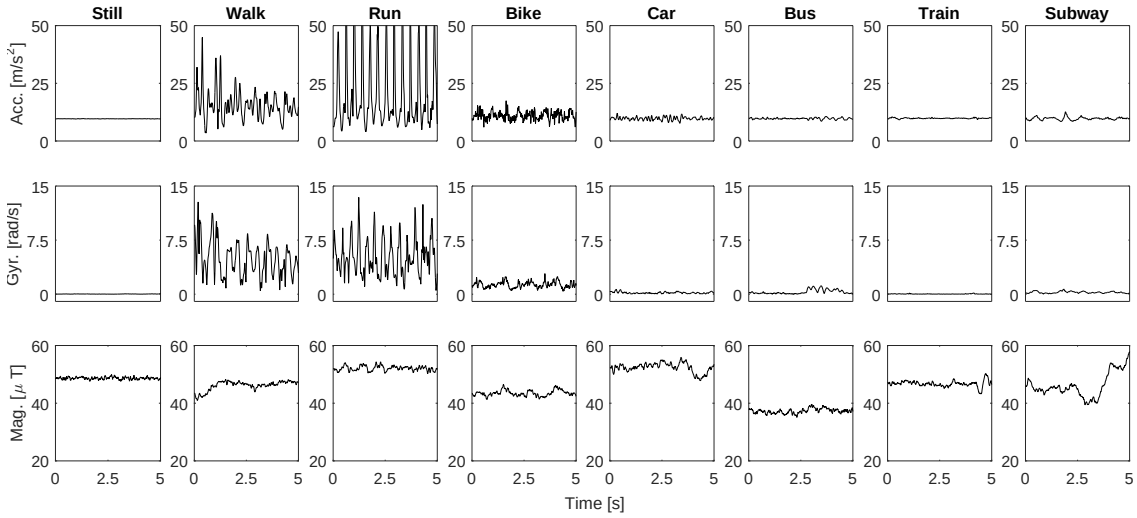
II. Dataset

The Sussex-Huawei Locomotion-Transportation (SHL) dataset is one of the biggest multimodal dataset for transportation and locomotion mode recognition from mobile devices [10], [32]. The dataset was recorded over 7 months by 3 users engaging in 8 different transportation modes: Still, Walk, Run, Bike, Car, Bus, Train and Subway. The duration of the dataset is 2812 hours, corresponding to a travel distance of 17,562 km in the south-east of the UK. The data was recorded using 4 smartphones placed at different locations on the body (hip pocket, hand, backpack, torso) and one body-worn unstabilized camera mounted on the chest and facing forwards (see Fig. 1). The dataset contains 16 sensor modalities including motion, sound and vision. The dataset was used as in the recent SHL challenge 2018: a competition on motion sensor-based transportation activity recognition [17], [18].

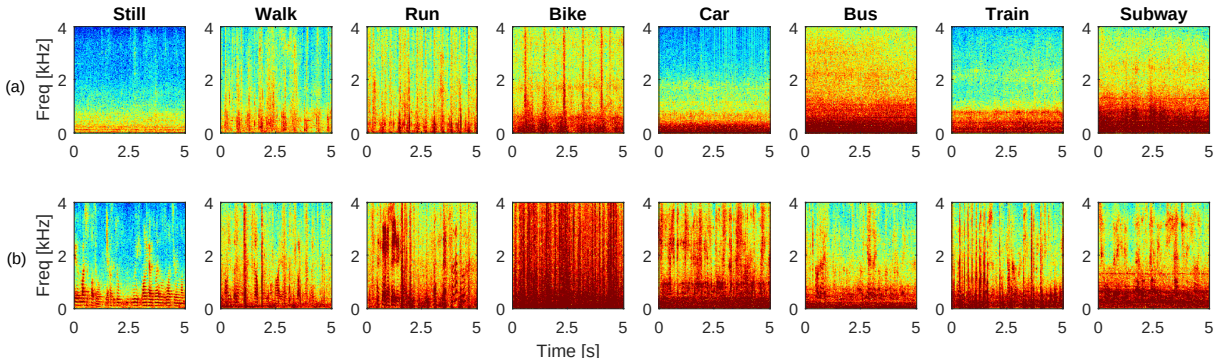
The motion, sound and vision data contained in the SHL dataset enables us to investigate the potential of data fusion for transportation mode recognition. For ease of comparison, we use exactly the same training and testing data partitioning scheme as in the SHL challenge 2018 [17]. Specifically, we use the multimodal sensor data recorded by the first participant with hand smartphone during 82 days (5-8 hours per day), which is partitioned in 62 days (271 hours) for training and 20 days (95 hours) for testing. Fig. 2 depicts the duration of each class activity in the training and testing datasets.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal



(a) Motion sensors for each transportation activity: Acceleration, Gyroscope and Magnetometer. The magnitude, as a combination of the data from the X, Y and Z axes, is displayed.



(b) Sound sensor: spectrogram of sound clips (5 seconds) for each transportation activity. The first row: clean sound during transportation. The second row: noisy sound (with environmental noise).

(c) Vision sensor: images from the body-worn camera for each transportation activity.

Fig. 3: Motion, sound and vision sample data in the SHL dataset.

The motion sensors include acceleration, gyroscope and magnetometer, which are all sampled at 100 Hz. The sound sensor (microphone) originally records sound at a sampling rate of 16 kHz, which is downsampled to 8 kHz before processing. The vision sensor (camera) takes one picture every 30 seconds (i.e. sampling rate 1/30 Hz).

A. Data interpretation

Fig. 3 visualizes exemplary data from the motion, sound and vision sensors in the SHL dataset.

Fig. 3(a) depicts the magnitude of the data provided by accelerometer, gyroscope and magnetometer, respectively. As a combination of the X/Y/Z-axes, the magnitude is robust to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

device orientation and rotation. Accelerometer shows higher energy for Walk, Run and Bike than for the other five activities. The accelerometer also shows evident cyclic behaviour for the Walk and Run activities. Similar observations can be made in the gyroscope data. The magnetometer does not seem to show visually distinctive patterns for different activities. This visual inspection shows that while some sensors provide clearly distinct signatures for some activities, distinguishing all 8 classes appears challenging, which motivates the use of machine learning methods capable of representation learning - such as deep learning - further on in this article.

Fig. 3(b) compares the short-time Fourier transform (STFT) spectrogram of the sound recorded during the 8 transportation activities. One big challenge for sound recognition is the influence of environmental noise. The first row shows a clean sound captured during transportation (without additional noise from the environment). The sound segment tends to show different spectrogram patterns for each activity. For instance, the activities Still, Car, Bus, Train and Subway tend to present different energy distribution in the low and high frequencies, while the activities Walk, Run and Bike tend to present different cyclic behaviour. In practice, the clean sound of each transportation activity is usually overlapped with additional noise from the environment, such as wind, friction, human speech, and other sound events nearby, as shown in the second row of Fig. 3(b). These environmental noises are typically much stronger than the clean transportation sound. This significantly increases the challenges when recognizing transportation activities.

Fig. 3(c) compares the 8 transportation modes taken by the front-facing camera. The resolution of the photos is 1024×576 pixels. In this example, most of the images are easy to recognize due to the relevant information provided by the environment. For example, for Bike we clearly see the handlebar with the hand on it; for Car we can see the roof and part of the dashboard of the car, as well as the road. The seats, bars, frame of the windows, shapes of the doors that appear in Bus, Train and Subway provide good-quality information to recognize them, although distinguishing between these three transportation modes becomes already more difficult. In Still, the user is inside and might be sitting on his couch looking at the room. Walk and Run are more challenging to distinguish as the movements of the arms and the places the user go could be very similar. Note that not all the pictures in the dataset are so nicely represented: some photos are tilted, blurred, rotated, upside-down, bright, dark or occluded due to the position, orientation, time of the day or movements of the user, which result in a more challenging task to recognize the transportation modes.

III. Mono-modal Classifiers

Fig. 4 illustrates the general processing pipeline of multimodal fusion. We first train three independent mono-modal classifiers with the motion, sound and vision data, respectively, and then fuse their results from better recognition performance.

All the three mono-modal classifiers are based on convolution neural networks (CNN). Each classifier predicts the probability (in the range [0, 1]) of each transportation activity, which is fed to the subsequent data fusion stage. The motion and sound classifiers process the sensor data per 5-second frames (one decision every 5 seconds) while the vision classifier processes every image (one decision every 30 seconds).

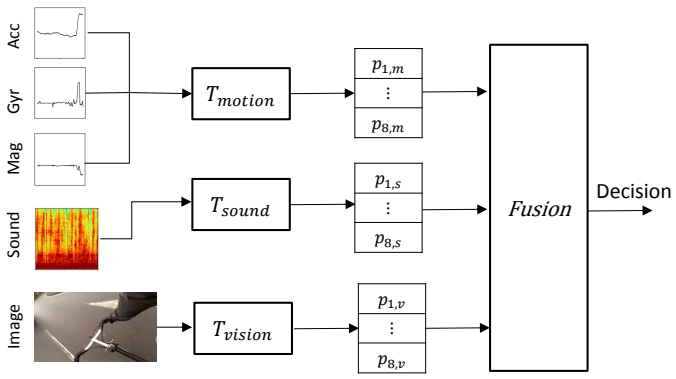


Fig. 4: General pipeline of multimodal fusion for transportation mode recognition. Each classifier T (m-motion, s-sound, and v-vision) predicts $p_{c,t} \in [0, 1]$, the probability of each of the eight class activities c, where $t \in \{m, s, v\}$ denotes the sensor modality. The outputs from the three classifiers are fused to make a joint decision on the mode of transportation.

A. Motion classifier

We have three motion sensors, i.e. accelerometer, gyroscope, magnetometer, each containing three channels of measurement along the X-, Y-, and Z-axis of device. Since the pose and orientation of the smartphone is unknown, we combine the three channels by computing the magnitude, i.e.

$$s_{acc}(i) = \sqrt{s^2_{acc\_x}(i) + s^2_{acc\_y}(i) + s^2_{acc\_z}(i)} \quad (1)$$

$$s_{gyr}(i) = \sqrt{s^2_{gyr\_x}(i) + s^2_{gyr\_y}(i) + s^2_{gyr\_z}(i)} \quad (2)$$

$$s_{mag}(i) = \sqrt{s^2_{mag\_x}(i) + s^2_{mag\_y}(i) + s^2_{mag\_z}(i)} \quad (3)$$

where i denotes the time index.

We convert the time-domain raw data to the frequency-domain, and then cascade the data from three sensors into a vector as

$$\boldsymbol{S}_F = \begin{bmatrix} \boldsymbol{S}_{acc} \\ \boldsymbol{S}_{gyr} \\ \boldsymbol{S}_{mag} \end{bmatrix}, \quad (4)$$

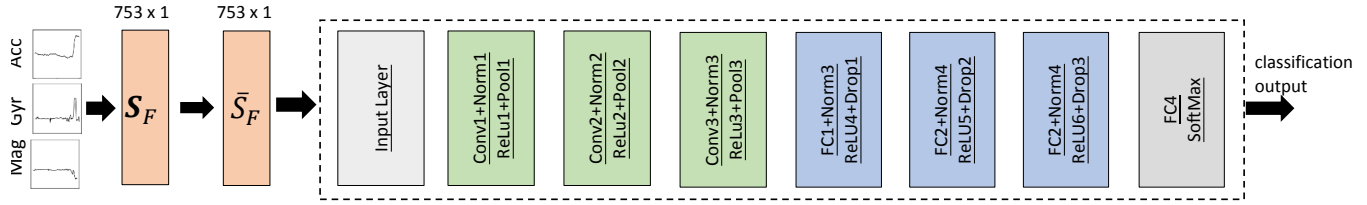
where $\boldsymbol{S}_{acc}/\boldsymbol{S}_{gyr}/\boldsymbol{S}_{mag}$ denotes the magnitude of the Fourier transform of $\boldsymbol{s}_{acc}/\boldsymbol{s}_{gyr}/\boldsymbol{s}_{mag}$ in one frame (retaining frequencies $[0, fs/2]$). Given the frame length 500, the size of $\boldsymbol{S}_{acc}/\boldsymbol{S}_{gyr}/\boldsymbol{S}_{mag}$ is 251×1, and therefore the size of $\boldsymbol{S}_F$ is 753×1.

The data $\boldsymbol{S}_F$ in each frame is normalized into the range $[0, 1]$ before classification, using

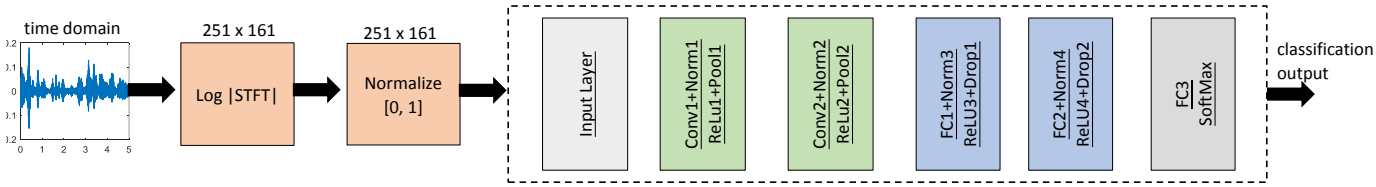
$$\bar{S}_F(k) \leftarrow \frac{S_F(k) - Q_5(k)}{Q_{95}(k) - Q_5(k)}, \quad (5)$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

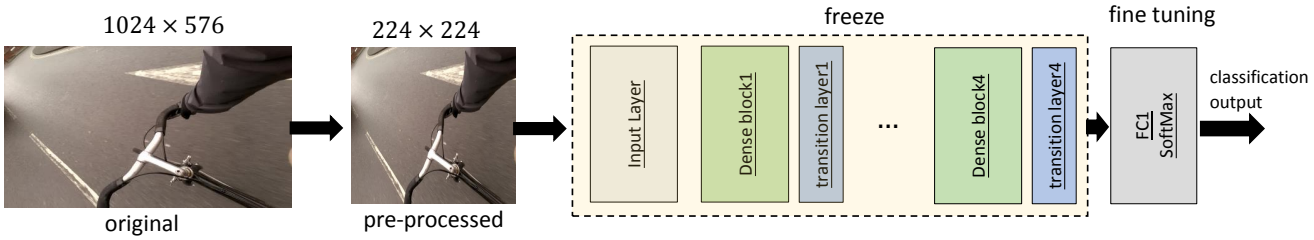
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal



(a) Motion classifier based on the convolutional neural network.



(b) Sound classifier based on the convolutional neural network.



(c) Vision classifier based on the adapted DenseNet169.

Fig. 5: Processing pipelines of the mono-modal classifiers.

TABLE I: Configuration of the convolutional neural network for motion-based classification corresponding to Fig. 5(a).

Input layer	size: (753, 1)
Conv1/Conv2/Conv3	number: 100; size: (15,1); stride: (1,1); padding: (0,0)
FC1/FC2/FC3	nodes: 300
Drop1/Drop2/Drop3	50%
FC4	nodes: 8
Norm1-6	mini-batch: 500

where $\bar{S}_F(k)$ denotes the k-th frequency bin in $\boldsymbol{s}_F$, $Q_{95}(k)$ and $Q_5(k)$ denote the quantile 95 and quantile 5, respectively, across all the frames in the training data.

Fig. 5(a) illustrates the deep architecture for motion-based classifier (T_{motion}), which we initially developed in [18]. The architecture consists of an input layer, multiple CNN and fully-connected neural network (FCNN) blocks and a decision block. The input layer receives and stores the frequency-domain motion sensor data $\boldsymbol{S}_F$. Each CNN block sequentially consists of a convolutional layer, a batch normalization (Norm) layer, and a nonlinear (ReLu) layer. Each FCNN block sequentially consists of a fully-connected (FC) layer, a batch normalization (Norm) layer, a nonlinear (ReLu) layer and a dropout layer. The decision block consists of a fully-connected layer, a nonlinear (Softmax) layer and a classification layer which infers the transportation mode. Table I gives the detailed configuration of the neural network.

For both training and testing dataset, we slide through the magnitude sensor data with a window of length 5 seconds and skip size of 5 seconds, generating framed data each containing 500 samples. This generates 195,688 frames of training data and 68,382 frames of testing data. The classification is conducted per individual frame. We use the Matlab Deep Learning Toolbox to implement the CNN classifier, using the stochastic gradient descent with momentum (SGDM) optimizer with default learning parameters.

B. Sound classifier

For sound data, we compute STFT spectrogram in each 5-second frame and then feed it to the classifier. The STFT spectrogram is computed with a sliding window of length 500 and half overlap. Therefore, the size of the spectrogram of the 5-second frame is 251×161. Let's represent the STFT in a frame as $S(k,l)$, where k and l denote the frequency and the STFT subframe indices, respectively.

To reduce the dynamic range of the data, we compute the log spectrogram as

$$A(k,l) = \log_{10}|S(k,l)|, \tag{6}$$

where $|\cdot|$ denotes the absolute value. We then normalize the data to the range of $[0,1]$ as

$$I(k,l) = \frac{A(k,l) - A_{min}}{A_{max} - A_{min}}, \tag{7}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

TABLE II: Configuration of the convolutional neural network for sound-based classification corresponding to Fig. 5(b).

Input layer	size: (251, 161)
Conv1/Conv2	number: 32; size: (5,5); stride: (1,1); padding: (0,0)
Pool1/Pool2	max pooling: (2,2); stride: (1,1); padding: (0,0)
FC1/FC2	nodes: 300
Drop1/Drop2	50%
FC3	nodes: 8
Norm1-4	mini-batch: 150

where A_{max} and A_{min} denote the maximum and the minimum values in the log spectrogram $A(:,:)$ throughout the training dataset.

Fig. 5(b) illustrates the deep architecture of the sound-based classifier (T_{sound}), which we initially developed in [24]. The convolutional neural network consists of an input layer, two CNN and two FCNN blocks, and an output decision block.

The input layer receives and stores the original spectrogram I. Each CNN block sequentially consists of a convolutional layer, a batch normalization (Norm) layer, a nonlinear (ReLu) layer and a pooling layer. Each FCNN block sequentially consists of a fully-connected (FC) layer, a batch normalization (Norm) layer, a nonlinear (ReLu) layer and a dropout layer. The decision block consists of an FC layer, a nonlinear (Softmax) layer which outputs the classification result. Table II gives the detailed configuration of the neural network.

We slide through the training dataset with a window of length 5 seconds and skip size of 20 seconds, generating data frames each containing 40,000 samples. This generates 65,240 frames of training data. For testing, we use a sliding window of length 5 seconds and skip size of 5 seconds. This results in 68,382 frames of testing data, which is the same size as the motion testing data. We use the Matlab Deep Learning Toolbox to implement the CNN classifier, using the stochastic gradient descent with momentum (SGDM) optimizer with default learning parameters.

C. Vision classifier

For image data, we employ a preprocessing procedure which resizes each image from 1024×576 to 224×224 before feeding it to the classifier.

Fig. 5(c) shows the vision classifier, as developed in [23], which is an adaptation of DenseNet169 [28]. DenseNet169 is a pre-trained CNN model on the ImageNet dataset for image recognition [29]. DenseNet169 consists of several dense blocks, transition layers and finally a classification layer. The dense blocks, each containing multiple densely connected convolution layers, are connected via a transitional layer, which consists of a convolution and a pooling layer. The dense blocks extract features from the input image, which are fed to the decision block for classification.

DenseNet169 was originally trained for image classification, and can not be used for transportation mode recognition directly. We employ a transfer learning scheme that adapts the DenseNet169 model to our classification problem. Specifically, we freeze the architecture and parameters of the DenseNet169 model except the 4th (and last) dense block. We replace the last FC layer and decision layer by an FC connected layer with 512 neurons followed by 8 Softmax activated neurons, which predicts the probability of each transportation activity. The parameters of the decision block are fine-tuned using the training data. Table III gives the detailed configuration of the neural network.

TABLE III: Configuration of the adapted DenseNet169 for vision-based classification corresponding to Fig. 5(c).

Input layer	size: (224, 224)
Dense block 1-3	default (frozen)
Transition 1-3	default (frozen)
Dense block 4	default (transfer learned)
FC1	nodes: 512 (fine-tuned)
FC2	nodes: 8 (fine-tuned)

Following the same training/testing data split scheme, we have 31,287 images in the training set and 10,781 images in the testing set. The vision classifier (T_{vision}) is implemented with the Python Keras library using TensorFlow as the backend computing library.

IV. Multimodal Sensor Fusion

The motion, sound and vision sensors work independently with different sampling rates. It is necessary to synchronize the data before fusing the classifier results. We first introduce the data synchronization method and then present the two data fusion schemes: ensemble decision and adaptive fusion. Finally, we employ a post-processing scheme that can further improve the recognition accuracy.

A. Synchronization

When recording sensor data, the smartphone and the camera both log the absolute world time (Unix Epoch time). Following the procedure described in the document[1] "Data organisation and file formats", we can retrieve the absolute world time for each frame of sound and motion data and for each image, as illustrated in Fig. 6.

The motion and sound classifiers make a decision per 5 seconds, while the vision classifier makes a decision per 30 seconds. We thus need to interpolate the vision classifier output to make it consistent with the output from the other two classifiers. We use the zero-order hold rule for interpolation [30]. Specifically, as exemplified in Fig. 6, the decision of the current image is retained for 30 seconds until the next image. The decisions of the motion, sound, vision classifiers are then fused at the same world clock time.

B. Ensemble decision

The ensemble decision scheme is based on simple mathematical rules to select the transportation mode (class) based on the outputs from multiple classifiers. Following the suggestions in [31], we consider the following four fixed rules: Majority Voting, Borda Count, Sum Rule and Product Rule. The first two are based on the output labels while the latter two are based on the class probability.

[1] http://www.shl-dataset.org/download/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

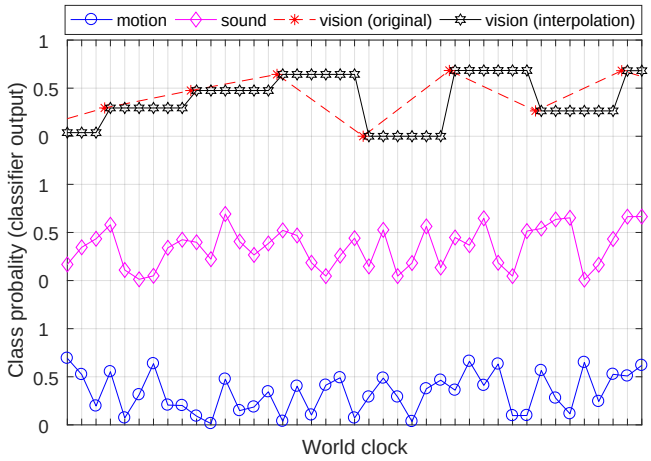


Fig. 6: Illustration of data synchronization. The motion and sound classifiers make a decision every 5 seconds (the space between two neighbouring ticks on the x-axis) while the vision classifier makes a decision every 30 seconds. The curves denote the probability of one class predicted by each classifier. The vision classifier output is interpolated with the zero-order hold rule.

Let us use $p_{c,t}$ to represent the predicted probability of the class c by the classifier t. Suppose we have C classes and T classifiers, i.e. $c \in [1, \cdots C]$ and $t \in [1, T]$. The decision of the classifier t would be

$$d_t = \underset{c\in[1,C]}{\operatorname{argmax}}\, p_{c,t}. \tag{8}$$

Majority Voting (MV) counts the class that appears the most in a sample, among the multiple classifiers. Let us use n_c denotes the occurrence of the class c, the majority voting rule is expressed as

$$d_{MV} = \underset{c\in[1,C]}{\operatorname{argmax}}\, n_c. \tag{9}$$

In **Borda Count** (BC), all the classes are ranked based on their predicted probability and are given weights based on the rank. For instance, the first one with the highest probability is given a weight $C-1$, the second one is given a weight $C-2$, and so forth, until the last one given a weight 0. We sum up the weights from all the classifiers and choose the one with the highest weight. In this way, the decision is less dependent on the probability value. Suppose the weight of class c by the classifier t is $w_{c,t}$, the decision is given by

$$W_c = \sum_{t=1}^{T} w_{c,t}, \tag{10}$$

$$d_{BC} = \underset{c\in[1,C]}{\operatorname{argmax}}\, W_c \tag{11}$$

The **Sum Rule** adds up the probabilities of each class across all classifiers and selects the one with the highest score as the transportation mode. This is expressed as

$$S_c = \sum_{t=1}^{T} p_{c,m}, \tag{12}$$

$$d_S = \underset{c\in[1,C]}{\operatorname{argmax}}\, S_c \tag{13}$$

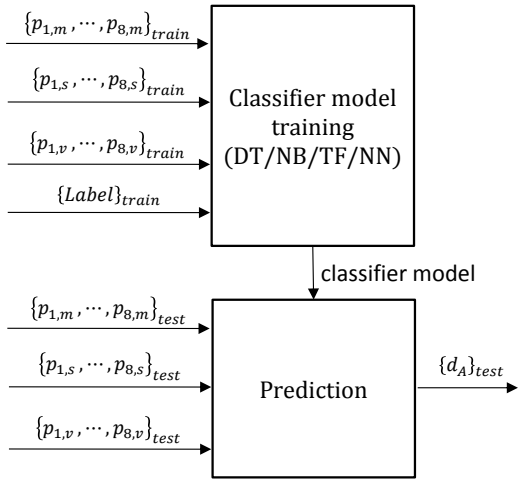


Fig. 7: Processing pipeline of the adaptive fusion scheme.

Product Rule is the same as Sum Rule except that it multiplies the probabilities of each class across all classifiers instead of adding them up. This is expressed as

$$P_c = \prod_{t=1}^{T} p_{c,m}, \tag{14}$$

$$d_P = \underset{c\in[1,C]}{\operatorname{argmax}}\, P_c \tag{15}$$

C. Adaptive fusion

In the adaptive fusion scheme, we try to learn the relationship between the outputs from multiple mono-modal classifiers and the joint decision with a classical machine learning classifier (adaptive fuser), such as naive Bayes (NB), decision tree (DT), random forest (RF) and multi-layer perceptron neural network (NN).

Fig. 7 depicts the processing pipeline of the adaptive fusion scheme. The input to the adaptive fuser is the set of class probabilities $\{p_{c,t}\}$ with $c = [1, \cdots, C]$, $t = [1, \cdots, T]$, and the output of the fuser is the joint decision $d_A \in [1, C]$.

The parameters of the fusing classifier model are obtained by feeding the mono-modal classifier outputs for the training data $\{p_{c,t}\}_{train}$ and the ground-truth label $\{Label\}_{train}$. The outputs $\{p_{c,t}\}_{train}$ is obtained via leave-one-out cross-validation. Specifically, we divide the training data into K-folds. For each fold, we train the mono-modal classifier with the K-1 folds and test with this fold. Cascading the testing results for all the K folds, we obtain the mono-modal classifier outputs for the whole training set, i.e. $\{p_{c,t}\}_{train}$. The mono-classifier output for the testing set $\{p_{c,t}\}_{test}$ is obtained by feeding the testing data to the classifier trained with the whole training set.

The adaptive fuser is implemented with the Python Scikit-learn library, using default parameters during training.

D. Post-processing

The classification system makes a decision every frame (5 seconds). Since the transportation mode of a user typically

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

continues for a certain period and there is a strong correlation between neighbouring frames [18], we reasonably assume that the transportation mode will remain unchanged for a certain time, e.g. in a window consisting of F frames. We employ a majority voting scheme to further improve the recognition performance at individual frames.

Suppose the prediction results in the f frame is $d(f)$ and the results in the previous $F-1$ frames is $d(f-F+1), \cdots, d(f-1)$. The occurrence of each activities in these F continuous frames is counted as $n_f(1), \cdots, n_f(C)$. The transportation mode of the current frame is determined as

$$\bar{d}_{post}(f) = \underset{c\in[1,C]}{\operatorname{argmax}}\, n_f(c). \tag{16}$$

In the experiment in Sec. V-D, we try different window length varying from 5 seconds (1 frame) to 180 seconds (36 frames) to see the impact of the window length on the recognition performance.

V. Experimental Results and Analysis

A. Evaluation Measure

We use F1-score over all the activities to evaluate the recognition performance using the testing dataset.

Let M_{ij} be the (i,j)-th element of the confusion matrix. It represents the number of samples originally belonging to class i which are recognized as class j. Let $C = 8$ be the number of classes. The F1-score is defined as below.

$$\text{recall}_i = \frac{M_{ii}}{\sum_{j=1}^{C} M_{ij}}, \tag{17}$$

$$\text{precision}_j = \frac{M_{jj}}{\sum_{i=1}^{C} M_{ij}}, \tag{18}$$

$$F1 = \frac{1}{C}\sum_{i=1}^{C} \frac{2\cdot \text{recall}_i \cdot \text{precision}_i}{\text{recall}_i + \text{precision}_i}, \tag{19}$$

We compare the F1-score achieved using one, two and three modalities, respectively. For two modalities, we consider different combinations, i.e. *ms* - {*motion, sound*}, *sv* - {*sound, vision*}, *mv* - {*motion, vision*}. For three modalities, we consider *msv* - {*motion, sound, vision*}.

B. Single modality

The F1-scores of each mono-modal classifier are given in Table IV as a baseline performance. Sound achieves the highest recognition performance (82.2%), followed by motion (79.4%), and vision achieves the lowest performance (72.8%). Fig. 8(a) shows the confusion matrices for these 3 modalities.

Sound is better at classifying the vehicle activities (Car, Bus, Train and Subway) than motion sensors. This is because each vehicle transportation typically emits unique sound that distinguishes itself from other activities, but presents similar motion patterns. Motion sensor is better at classifying pedestrian activities (Still, Walk, Run, Bike) than sound. This is because pedestrian and biking activities require strong user engagement, but emit sound which is much weaker than environmental noise. This implies that the combination of the two modalities potentially leads to better recognition result. Vision performs poorly at distinguishing between Still, Walk and Run, possibly due to the operating environment of the three activities are similar. Vision performs relatively better at distinguishing the remaining five activities. Vision performs better at distinguish vehicle activities than motion, but worse than sound. However, vision performs the best when identifying the Subway activity. Some objects, such as people and seats, can be used to effectively infer the external environments. Overall, the recognition results using motion and using sound are truly complementary. Additionally using vision could further improve the discriminability between vehicle activities.

TABLE IV: F1-score of mono-modal classifier.

Modality	*motion*	*sound*	*vision*
F1 [%]	79.4	82.1	72.8

TABLE V: F1-score of multimodal classifier.

		F1 score [%]			
	Method	*m+s*	*s+v*	*m+v*	*m+v+s*
Ensemble Decision	Majority Voting	79.4	79.4	82.2	89.9
	Borda Count	87.9	81.7	82.3	88.4
	Sum	89.8	90.0	88.3	93.0
	Product	**91.5**	**91.0**	**89.2**	**94.5**
Adaptive Fusion	Naive Bayes	89.0	88.4	86.4	92.3
	Decision Tree	87.7	85.2	84.5	90.7
	Random Forest	**92.5**	**90.9**	**90.0**	**95.5**
	Neural Network	90.7	90.4	88.3	94.6

C. Multimodality

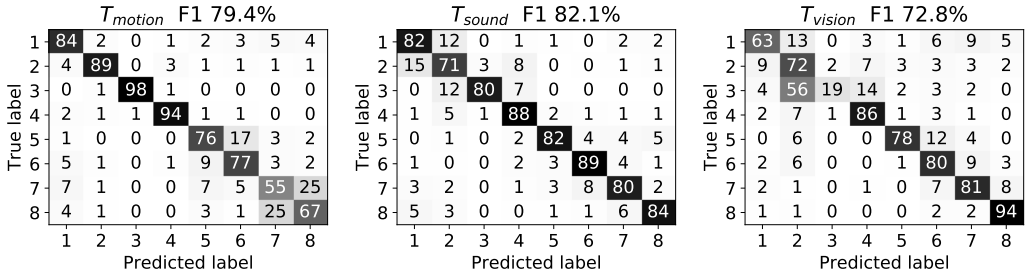
Table V compares the data fusion results applied to all the possible combinations of the three sensor modalities.

For ensemble decision, the two probability-based approaches (Sum and Product) significantly outperform the two label-based approaches (MV and BC). When combining the three modalities, the highest F1-score achieved by the probability based and the ensemble decision based approaches are 94.5% (Product) and 89.9% (MV), respectively. For the two probability-based approach, the Product Rule (94.5%) performs slightly better than the Sum Rule (93.0%). When combining two modalities, the label-based fusion approach does not show evident advantages over using single modality while the probability-based approaches achieve higher F1-scores. This is possibly due to a limited amount of classifiers available (maximum three) for data fusion. The probability-based approaches perform more robustly when only a few classifiers are available. Furthermore, the label-based fusion approaches loose information by operating on a crisp decision, whereas probability-based approaches retain more information which can be exploited during fusion.

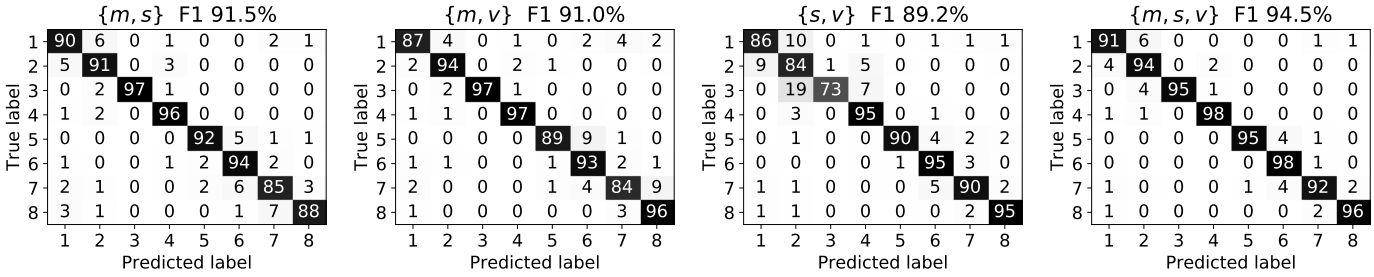
For adaptive fusion, random forest performs the best among the four fusers. When combining the three modalities, the four fusers achieve F1-scores of 92.3% (NB), 90.7% (DT), 95.5% (RF) and 94.6% (NN), respectively. In comparison, the ensemble decision method achieves the highest F1-score

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

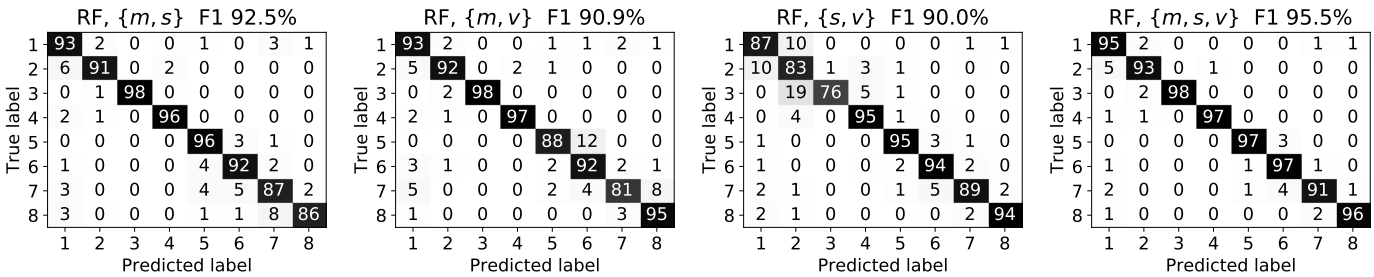
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal



(a) Baseline performance of mono-modal classifiers.



(b) Ensemble decision: Product Rule



(c) Adaptive fusion: Random Forest

Fig. 8: Confusion matrices for (a) mono-modality; (b) multimodal data fusion using Product Rule; (c) multimodal data fusion with adaptive fuser Random Forest. The eight class activities: 1-Still, 2-Walk, 3-Run, 4-Bike, 5-Car, 6-Bus, 7-Train, 8-Subway.

of 94.5% (Product), which is about 1% lower than the RF. When combining two modalities, the Product Rule and the RF achieve F1-scores of 91.5% vs 92.5% for {*motion, sound*}, 91.0% vs 90.9% for {*sound, vision*}, and 89.2% vs 90.0% for {*motion, vision*}, respectively. Overall, the adaptive fusion method performs slightly better than the ensemble decision method. The increase of performance by adaptive fusion can be justified by the capacity of machine learning classifiers identifying specific relationships between the classifier outputs and the joint decision. However, the downside is that in adaptive fusion the classifier might over-fit on the training data and thus may not generalize well to unseen data.

Fig. 8 show the confusion matrices obtained by the different data fusion strategies. We consider Product and RF for ensemble decision and adaptive fusion, respectively. As suggested in Sec. V-B and also confirmed in Table V, data fusion can improve the recognition performance significantly by exploiting the complementarity between motion, sound and vision. For instance, {*sound, motion*} improves the recognition performance of each class activity over using either sound or motion alone. Similar observations can be made for {*motion, vision*} and {*sound, vision*}.

For ease of comparison, we extract the diagonal elements in each confusion matrices and depict them in Fig. 9. The diagonal element indicates the ability of the classifier to identify the corresponding class. We only consider the Product Rule for data fusion. For single modality, motion performs the best at identifying Still, Walk, Run and Bike; sound performs the best at identifying Bus and Car; vision performs the best at identifying Train and Subway. For dual modality, {*motion, sound*} performs the best at identifying Still, Walk, Run; and performs equally well as other dual modalities at identifying Bus and Car; and worse at identifying Train and Subway. {*Motion, vision*} performs the best at identifying Walk, Run, Bike, and worst at Still, Car, Bus and Train. {*Sound, vision*} performs the best at identifying vehicles, including Train, Subway, Bus and Car; and performs worst at identifying Still, Walk, Run and Bike. Finally, for triple-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

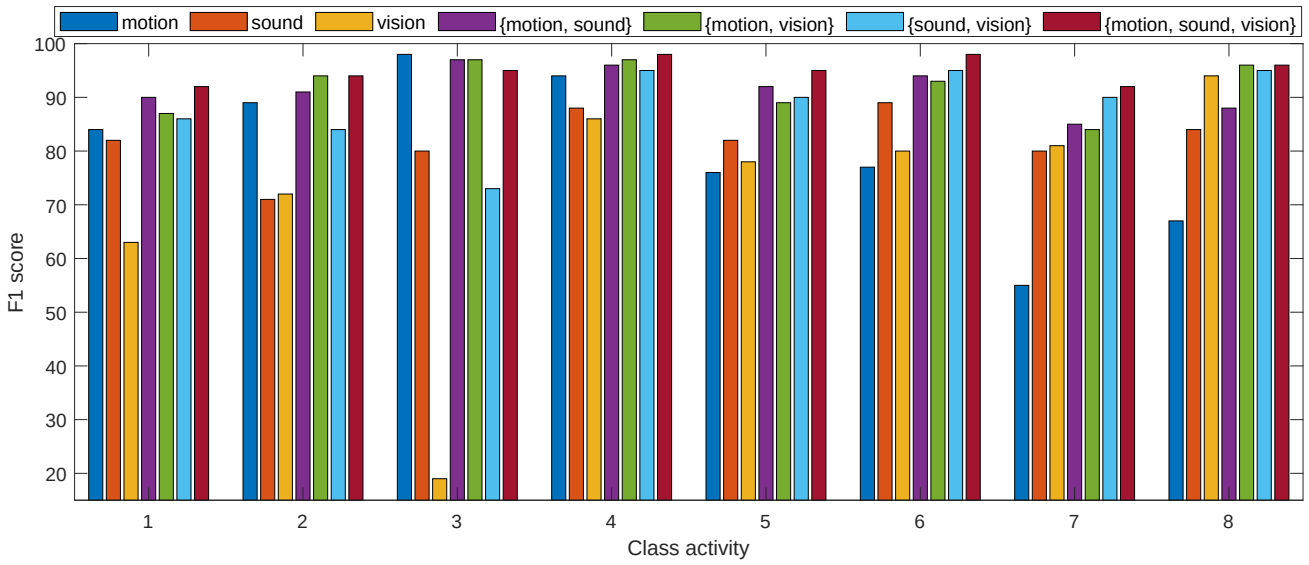


Fig. 9: The recognition performance for each individual class activity by combining different sensor modalities. The eight class activities: 1-Still, 2-Walk, 3-Run, 4-Bike, 5-Car, 6-Bus, 7-Train, 8-Subway.

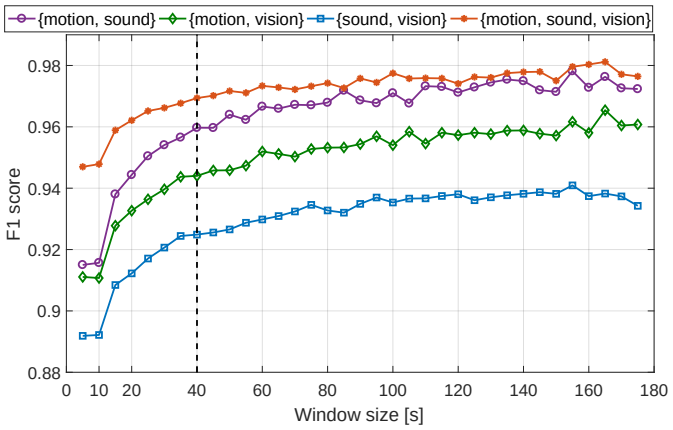


Fig. 10: Post-processing results (F1 score) on the multimodal classifier with various combination of sensors. The smoothing window size varies from 5 seconds to 180 seconds at a step of 5 seconds.

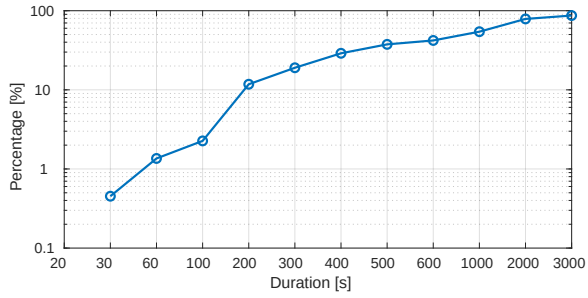


Fig. 11: Cumulative distribution of the duration of each continuous activity period in the testing dataset.

modality, the performance can be improved for identifying each class activity over using dual modality.

D. Post-processing results

We apply post-processing to the multimodal classifier (Product Rule) with various combination of sensors: $\{motion, sound\}$, $\{motion, vision\}$, $\{sound, vision\}$ and $\{motion, sound, vison\}$. Fig. 10 depicts the post-processing results achieved with a smoothing window sizing from 5 seconds (1 frame) to 180 seconds (36 frames) at a step of 5 seconds.

For each multimodal classifier, the post-processing performance shows a similar variation trend with increasing window size. The F1 score improves remarkably (e.g. from 91.5% to 94.4% for $\{motion, sound\}$) when the smoothing window size grows from 5 seconds to 15 seconds. The performance then improves quickly (e.g. from 94.4% to 96.0% for $\{motion, sound\}$) for smoothing window size $[15, 40]$ seconds, and then slowly (e.g. from 96.0% to 96.8%) for smoothing window size $[40, 80]$ seconds. The improvement becomes marginal when the smoothing window size is larger than 80 seconds; and then the improvement appears unstable when the window size is larger than 150 seconds.

This is explained by Fig. 11, which shows the duration of each continuous activity in the testing dataset. In Fig. 11, the y-axis denotes the cumulative distribution, i.e. the percentage ratio between the number of continuous activity periods with duration less than a certain value and the total number continuous activity periods. It can be observed that only 2% activities last less than 60 seconds, and 90% of activities last more than 200 seconds. A similar distribution can be observed in the training dataset (which is not shown here). This verifies the feasibility of temporal smoothing, and also indicates that it appears reasonable to choose a smoothing window of length around 40-60 seconds.

While the recognition performance improves with the smoothing window size, in a mobile computing scenario, the choice of post-processing window size will need to be

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

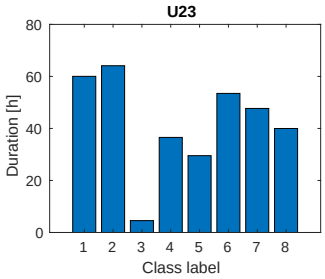


Fig. 12: The duration of each class activity in the new testing data U23, which contains the all data from the User 2 and 3 in the SHL dataset. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

decided based on the needs of the application. For applications requiring real-time response, a shorter post-processing window will be desired, while for applications doing longitudinal statistics where high accuracy is preferable a longer post-processing window should be employed.

VI. Generalization to unseen data

In Sec. V, the mono-modal classifiers and the adaptive fusers are trained and tested using different folds of data from the same user (User 1 - U1). In this section, we further investigate the generality of the proposed method with a new dataset (U23), which contains the data from User 2 and User 3 in the SHL dataset. The data collection protocol of U23 was the same as U1: using a smartphone at the hand position and a body-worn camera. The total duration of the data in U23 is 356 hours, with the duration of each class activity shown in Fig. 12.

We perform mono-modal classification with the three modalities, and perform multimodal fusion with Product-based ensemble decision and the RF-based adaptive fusion. We apply the same mono-modal classifiers and adaptive fusers, that are trained with U1 in Sec. V, to U23. This means the training and testing are conducted with the data from different users, which is a more challenging task as the three users in the SHL dataset tend to have different behaviours and habits and device wearing styles. This allows to evaluate the performance of the algorithms on an 'unseen' dataset.

Fig. 13 depicts the recognition and fusion result in terms of F1 score and confusion matrix. Comparing Fig. 8 (for U1) and Fig. 13 (for U23), it can be observed that the recognition performance (F1 score) of the three mono-modal classifiers drops significantly when training and testing with different users. Specifically, the performance of motion classifier drops 16.1pp (percentage point) from 79.4% to 63.3%; the sound classifier drops 12.6pp from 82.1% to 69.5%; and the vision classifier drops 26.8pp from 72.8% to 46.0%. Such behaviour is expected as we evaluate a user-specific model to unseen users.[2] Among the three modalities, the sound modality is the most robust to the variation of users as the microphone captures the sound from surrounding environment, which is not affected by the user behaviour. The motion modality is less robust than sound, as the behaviour varies with users. The vision modality is least robust to user variation, possibly because of the different styles in which users carried the body-worn camera.

The recognition performance (F1 score) of the multimodal fusion also drops. For instance, the Product-based tri-modal classifier $\{$*Motion, sound, vision*$\}$ drops 9.6pp from 94.5% to 84.9%, and the RF-based tri-modal classifier drops 12.3pp from 95.5% to 83.2%. Nevertheless, the benefit of fusion and the complementarity of the three modalities can still be observed from the confusion matrices in Fig. 8. For instance, motion is good at distinguishing between pedestrian activities and poor at vehicle activities; sound is good at distinguishing between vehicle activities and poor at pedestrian activities. It can also be observed that, in this experiment, motion performs poorly at identifying the Bike activity while sound and vision both perform better. Taking advantage of this complementarity, multimodal fusion always improves the performance over mono-modal classifier. For instance, the Product-based tri-modal classifier improves the performance of the sound classifier (the best performing mono-modal classifier) by 15.4pp from 69.5% (sound) to 84.9%. More importantly, this improvement (15.4pp for U23) is even higher than what we achieved for U1 (12.5pp from 82.1% to 94.5%, in Fig. 8). This implies that the multimodal fusion can improve the robustness to user variation.

Both RF-based and Product-based fusion schemes work effectively improving the performance over mono-modal classifiers. However, the RF-based scheme is less robust to user variation than the Product-based scheme. For instance, the RF-based tri-modal fusion (83.2%) performs 1.7pp lower than the Product-based fusion (84.9%) for U23 (in Fig. 13). This is in contrast to the result reported for U1 (in Fig. 8), where the RF-based scheme is 1pp higher than the Product-based scheme.

If we apply post-filtering, the recognition performance can be further improved. For instance, while the details are not reported in the paper, we observe that the F1 score of the Product-based tri-modal classifier is improved by 4.8pp, from 84.9% to 89.7%, with a smoothing window length 45 seconds.

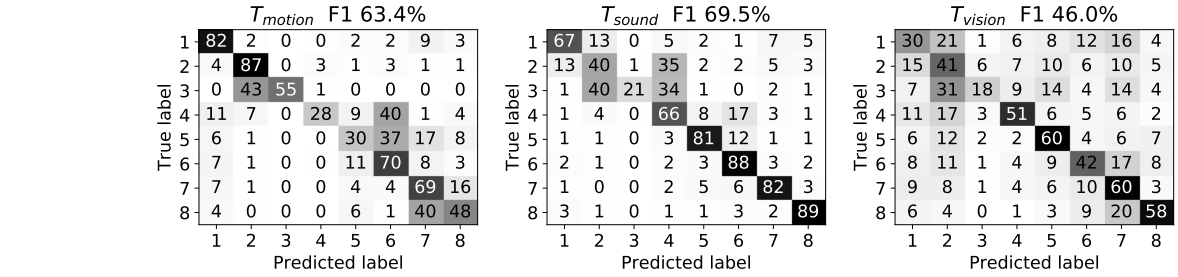
In short, the experimental results above verify the generality of the proposed multimodal fusion methods. While the recognition performance mono-modal classifiers drops due to user variation, the complementary of the three modalities can still be observed, and the fusion of any two or three modalities always improves the recognition performance over mono-modal classifier, and it also improves the robustness to user variation. Both Product-based and RF-based fusion scheme works well for multimodal fusion although the RF-based scheme shows slight performance drop due to user variation.

We would like to highlight that the experiment in this section mainly aims to assess the generality of the proposed method, rather than developing a user-independent recognition system, which is in practice should be trained on multiple

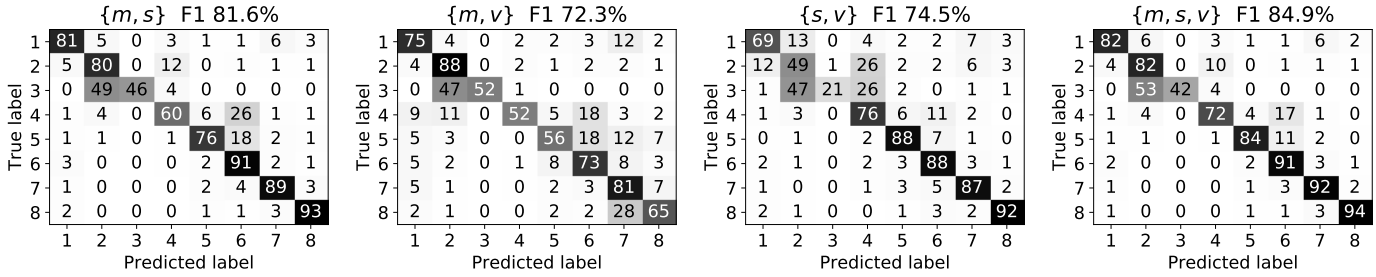
[2]Note that in the field of activity recognition, when a system is designed to generalize to new users a "leave-one-user-out" cross-validation is employed. In this article we test a user-specific model on new users as a way to emulate application to a new dataset. We employ U2 and U3 of SHL to emulate this 'new' dataset, as there are no other multimodal transportation datasets suitable for this analysis, to our knowledge.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

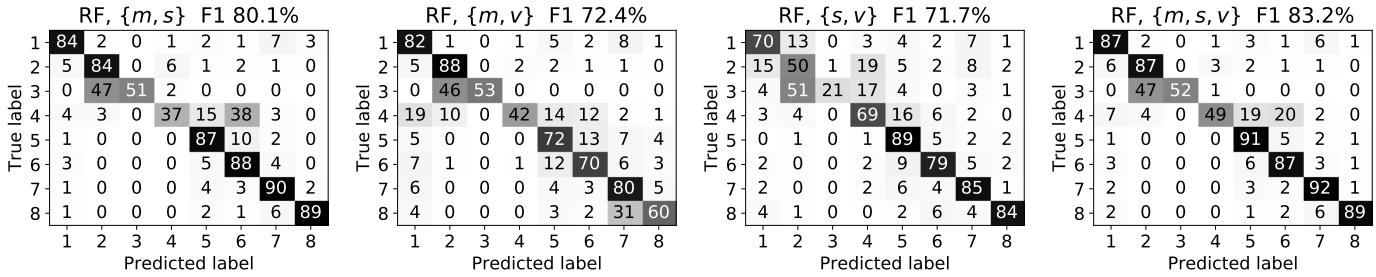
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal



(a) Baseline performance of mono-modal classifiers.



(b) Ensemble decision: Product Rule



(c) Adaptive fusion: Random Forest

Fig. 13: Confusion matrices for (a) mono-modality; (b) multimodal data fusion using Product Rule; (c) multimodal data fusion with adaptive fuser Random Forest. The mon-modal classifiers and the adaptive fuser are trained on User 1 and tested on User 2 and 3. The eight class activities: 1-Still, 2-Walk, 3-Run, 4-Bike, 5-Car, 6-Bus, 7-Train, 8-Subway.

users to capture the user variation. However this is the best evaluation of generalization to new data that we can perform using the SHL dataset (with only 3 users), and taking into account there are no other suitable public multimodal transportation datasets available for a similar analysis.

VII. Discussion

The mono-modal classifiers employed in this paper are adapted directly from our previous work [18], [23], [24], which are comparable to the state of the art. The sound and vision classifiers are among the first works that are applied to transportation mode recognition. The motion classifier was used to benchmark the SHL Challenge 2018 [17] and performed slightly worse than the winner of that challenge [32]. Since this paper mainly focuses on multimodal fusion, we did not aim to maximize the performance of each mono-modal classifier. However, we believe that, with the late integration strategy, the performance of the multimodal classifier would be further improved when each mono-modal classifier would be optimized independently. There are two interesting directions that could be investigated.

- First, all the three mono-modal classifiers employ a convolutional neural network. A recurrent neural network (e.g. LSTM [33]) could be employed to exploit the temporal correlation and to improve the recognition performance for time-series signals such as motion and sound. In principle a recurrent network could also be applied to video streams, however the images in the SHL dataset come from a timelapse camera which took a picture every 30 seconds.
- Second, over-fitting is a crucial issue in activity recognition, such as a classifier trained on one user with specific senor placement tends to show degraded performance for other users and sensor placement [10]. In this paper, the mono-modal classifiers simply employ a generic technique, e.g. dropout [34], to tackle the over-fitting problem. While the classifiers show promising

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

results on the same user, the performance drops significantly in case of user variation. In future, more techniques could be employed to tackle the over-fitting problem, including some techniques reported in the SHL challenge, such as augmented learning, transfer learning, and designing hand-crafted features [32], [35], [36].

In essence, multimodal fusion improves the recognition performance at the expense of increasing the sensor channels and also the computational complexity. For instance, we use a computer equipped with an Intel i7-4770 4-core CPU @ 3.40 GHz with 32 GB memory, and a GeForce GTX 1080 Ti GPU with 3584 CUDA cores @ 1.58 GHz and 11 GB memory; and the computation time of applying each mono-modal classifier on the testing dataset of User 1 is 7.5 seconds, 70.3 seconds, and 196.9 seconds, for motion, sound and vision, respectively. The vision classifier has the largest computational complexity, followed by sound and motion. The sound classifier has higher computational complexity than the motion classifier, as the sampling rate of sound (8k Hz) is much higher than motion (100 Hz). Current and upcoming smartphones offer increasingly powerful hardware acceleration for inference, which makes even seemingly complex models suitable for embedded execution (e.g. [37]). While a mobile phone implementation would show different numbers, the relative complexity of the different modalities is likely to be similar.

In this paper we focused on a modular fusion approach of "late integration" which allows to modularly combine classifiers together. Future work may explore the dynamic selection of classifiers to fuse at any given time to achieve particular application requirements, such as maximizing performance overall or for a particular set of activities, or minimizing power consumption. While a tri-model system performs the best, for dual-modality systems, $\{$*sound, motion*$\}$ achieves robust performance in most cases. For battery consumption optimization and low powered devices, we recommend to use in first instance the motion sensors, then combine it with sound sensor and finally with vision sensor. Also, for the specific task of recognizing the transportation mode, an efficient solution would be to combine all the three modalities but prioritizing motion by limiting the use of vision and sound. For instance, let us assume that motion could classify Still, Walk, Run, Bike and Vehicle. If the class is Vehicle, then the sound and vision can be used to further classify Car, Bus, Train and Subway. We can also activate modalities in specific situations. For instance, if in Bus, we only activate walk detection based on motion.

VIII. Conclusion

We applied data fusion methods to combine the output of three expert classifiers, dealing respectively with motion, sound and vision data, in order to improve the recognition of eight different transportation modes. Two sets of fusion techniques, ensemble decision (Majority Voting, Borda Count, Sum Rule and Product Rule) and adaptive fusion (Naive Bayesian, Decision Tree, Random Tree, Neural Network), are considered. Experimental results demonstrate that, by fusing any two modalities or all the three modalities, better recognition performance can always be achieved over using a single modality. The proposed multimodal fusion methods show good generality and can improve the robustness to user variation.

If we only look at the recognition result on the testing dataset of User 1, the best performance achieved by using a single modality is achieved by sound (F1 82.1%). When fusing three modalities, the best performance is achieved by the Product Rule (F1 94.5%) for ensemble decision, and is achieved by Random Forest (F1 95.5%) for adaptive fusing. This improves the recognition performance by 12.3pp (percentage point) and 13.3pp over the best mono-modal classifier (sound), respectively. The adaptive fuser improves performance by 1pp in the best case $\{motion, sound, vision\}$ compared to ensemble decision, and at worst led to only a minor decline in performance (-0.1pp with $\{sound, vision\}$). This indicates that, as an overall recommendation, an adaptive fuser should be favoured in the majority of the cases. In addition, the effect of post-processing method to recognize the transportation mode over a longer period of time already improves the F1 score by 2 pp within a 15-second window and 4 pp within a 45-second window. By comparing multiple combination of the modalities, i.e. $\{$*motion, sound*$\}$, $\{$*sound, vision*$\}$, $\{$*motion, vision*$\}$ and $\{$*motion, sound, vision*$\}$, we deducted that dual-modality based systems should prioritize motion and sound for more robustness and power-consumption efficiency.

Although the adaptive fuser performs better than ensemble decision for multimodal fusion, future development should consider the generalization of our methods on external datasets, as over-fitting might have occurred due to the sound and vision data closely related to the environment of the country where the dataset was collected (United Kingdom). However, to date no other dataset exists for such analysis to our knowledge.

More important than demonstrating a particular numerical value of the performance increase through fusion, this work opens up a space for the design of activity aware systems on smartphones which are able to dynamically balance power and performance requirements according to the needs of an application. Thanks to the modular "late integration" fusion approach which we follow here, the number of modalities and classifiers which are combined can easily be modulated. Exploring such a dynamic fusion remains the object of future work. Also, we might consider comparing the results of this research with a more complex classifier that should take as input directly all the three modalities. However, the improvements achieved here with the presented fusion methods are already outperforming mono-modality based classifiers, even though a more complex approach would be more desirable.

References

[1] J. Engelbrecht, M. J. Booysen, G. van Rooyen, and F. J. Bruwer, "Survey of smartphone-based sensing in vehicles for intelligent transportation system applications," *IET Intelligent Transport Sys.*, vol. 9, no. 10, pp. 924-935, 2015.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

[2] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62-74, 2017.

[3] E. Anagnostopoulou, J. Urbancic, E. Bothos, B. Magoutas, L. Bradesko, J. Schrammel, and G. Mentzas, "From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation," *J. Intelligent Information Sys.*, vol. 2018, pp. 1-22, 2018.

[4] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. IEEE Conf. Intelligent Transportation Sys.*, 2011, pp. 1609-1615.

[5] W. Brazil and B. Caulfield, "Does green make a difference: The potential role of smartphone technology in transport behaviour," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 93-101, 2013.

[6] J. Froehlich, T. Dillahunt, P. Klasnja, J. Mankoff, S. Consolvo, B. Harrison, and J. A. Landay, "Ubigreen: Investigating a mobile tool for tracking and supporting green transportation habits," in *Proc. SIGCHI Conf. Human Factors Computing Sys.*, 2009, pp. 1043-1052.

[7] C. Cottrill, F. Pereira, F. Zhao, I. Dias, H. Lim, M. Ben-Akiva, and P. Zegras, "Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore," *Transportation Research Record: J. Transportation Research Board*, vol. 2354, no. 1, pp. 59-67, 2013.

[8] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1321-1330, 2015.

[9] G. Castignani, T. Derrmann, R. Frank, and T. Engel, "Driver behavior profiling using smartphones: A low-cost platform for driver monitoring," *IEEE Intelligent Transportation Sys. Mag.*, vol. 7, no. 1, pp. 91-102, 2015.

[10] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen, "Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset," *IEEE Access*, vol. 7, pp. 10870-10891, 2019.

[11] P. Siirtola and J. Roning, "Recognizing human activities user independently on smartphones based on accelerometer data," *Int. J. Interactive Multimedia Artificial Intelligence*, vol. 1, no. 5, pp. 38-45, 2012.

[12] T. Feng and H. J. Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 118-130, 2013.

[13] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *Proc. ACM Conf. Embedded Networked Sensor Sys.*, 2013, pp. 1-14.

[14] H. Xia, Y. Qiao, J. Jian, and Y. Chang, "Using smart phone sensors to detect transportation modes," *Sensors*, vol. 14, pp. 20843-20865, 2014.

[15] X. Su, H. Caceres, H. Tong, and Q. He, "Online travel mode identification using smartphones with battery saving considerations," *IEEE Trans. Intelligent Transportation Sys.*, vol. 17, no. 10, pp. 2921-2934, 2016.

[16] S. Fang, Y. Fei, Z. Xu, and Y. Tsao, "Learning transportation modes from smartphone sensors based on deep neural network," *IEEE Sensors J.*, vol. 17, no. 18, pp. 6111-6118, 2017.

[17] L. Wang, H. Gjoreskia, K. Murao, T. Okita, and D. Roggen, "Summary of the Sussex-Huawei locomotion-transportation recognition challenge," in *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2018, pp. 1521-1530.

[18] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen, "Benchmarking the SHL recognition challenge with classical and deep-learning pipelines," in *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2018, pp. 1626-1635.

[19] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733-1746, 2015.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: An overview of DCASE 2017 challenge entries," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 411-415.

[21] P. Wang and A. F. Smeaton, "Using visual lifelogs to automatically characterize everyday activities," *Information Sciences*, vol. 230, no. 5, pp. 147-161, 2013.

[22] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia, "Privacy behaviors of lifeloggers using wearable cameras," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Computing*, 2014, pp. 571-582.

[23] S. Richoz, M. Ciliberto, L. Wang, P. Birch, H. Gjoreski, A. Perez-Uribe, and D. Roggen, "Human and machine recognition of transportation modes from body-worn camera images," in *Proc. Joint 8th Int. Conf. Informatics, Electronics & Vision and 3rd Int. Conf. Imaging, Vision & Pattern Recognition*, 2019, pp. 67-72.

[24] L. Wang and D. Roggen, "Sound-based transportation mode recognition with smartphones," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 930-934.

[25] S. Lee, J. Lee, and K. Lee, "Vehiclesense: A reliable sound-based transportation mode recognition system for smartphones," in *Proc. IEEE Int. Symp. A World of Wireless, Mobile Multimedia Networks*, 2017, pp. 1-9.

[26] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proc. 8th ACM Conf. Embedded Networked Sensor Sys.*, 2010, pp. 71-84.

[27] H. Gjoreski, M. Ciliberto, L. Wang, F. J. O. Morales, S. Mekki, S. Valentin, and D. Roggen, "The University of Sussex-Huawei locomotion and transportation dataset for multimodal analytics with mobile devices," *IEEE Access*, vol. 6, pp. 42592-42604, 2018.

[28] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2017, pp. 4700-4708.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. Li, "ImageNet large scale visual recognition challenge," *Int. J. Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.

[30] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, USA, 1989.

[31] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Sys. Mag.*, vol. 6, no. 3, pp. 21-45, 2006.

[32] M. Gjoreski, V. Janko, N. Rescic, et al., "Applying multiple knowledge to Sussex-Huawei locomotion challenge," in *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2018, pp. 1488-1496.

[33] F. J. O Morales and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, pp. 115, 2016.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[35] P. Widhalm, M. Leodolter, and N. Brandle, "Top in the lab, flop in the field? Evaluation of a sensor-based travel activity classifier with the SHL dataset," in *Proc. 2018 ACM Int. Joint Conf. 2018 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2018, pp. 1479-1487.

[36] V. Janko, M. Gjoreski, C. M. De Masi, et al., "Cross-location transfer learning for the sussex-huawei locomotion recognition challenge," in *Proc. 2019 ACM Int. Joint Conf. 2019 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2019, pp. 730-735.

[37] S. Richoz, A. Perez-Uribe, P. Birch, and D. Roggen, "Benchmarking deep classifiers on mobile devices for vision-based transportation recognition," in *Proc. 2019 ACM Int. Joint Conf. 2019 Int. Symp. Pervasive Ubiquitous Computing Wearable Computers*, 2019, pp. 803-807.

[38] P. Zappi, D. Roggen, E. Farella, G. Troster, and L. Benini, "Network-level power-performance trade-off in wearable activity recognition: A dynamic sensor selection approach," *ACM Trans. Embedded Computing Sys.*, vol. 11, no. 3, pp. 1-30, 2012.

[39] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98-125, 2017.

[40] J. H. Choi, and J. S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259-270, 2019.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.2987306, IEEE Sensors Journal

Sebastien Richoz is a Research Assistant in the School of Engineering and Informatics at the University of Sussex. He has research interests in computer vision and artificial intelligence. He acquired his Bachelor (2017, Yverdon) in Computer Science, major in Software Engineering and his Master (2019, Lausanne) in Engineering, major in Information and Communication Technologies, both from the University of Applied Sciences HES-SO in Switzerland. Then he joined the University of Sussex to work on human activity recognition and cancer research using machine learning techniques, image processing and object detection methods.

Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003; and the Ph.D degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he has been an Alexander von Humboldt Fellow in University of Oldenburg, Germany. From 2014 to 2017, he has been a postdoctoral researcher in Queen Mary University of London, UK. From 2017 to 2018, he has been a postdoctoral researcher in the University of Sussex, UK. Since September 2018, he has been a Lecturer in Queen Mary University of London. His research interests include audio-visual signal processing, machine learning, and robotic perception.

Philip Birch is a Reader in Engineering in the School of Engineering and Informatics at the University of Sussex. He has a research interest in computer vision and electro-optics. He graduated with a degree in Physics from the University of Durham in 1994. He gained a PhD from the same institution in 2000 in the field of adaptive optics. He worked from 2004 until 2007 commercially developing microscopic lithographic systems and three dimensional optical metrology systems within industry. Since then he has worked as a Research Fellow and then as Research Faculty investigating computer generated holograms, optical correlation for signal processing, computer vision object tracking and image processing.

Daniel Roggen (M'04) is Professor in Sensor Technologies at the University of Sussex, where he leads the Wearable Technologies Lab and directs the Sensor Technology Research Centre. His research focuses on computational behaviour analytics: the use of machine learning techniques, miniature intelligent sensor systems and other data sources to recognize, qualify, quantify and eventually understand human behaviours and the wider context in which they occur. He has established a number of recognized datasets for human activity recognition from wearable sensors, in particular the OPPORTUNITY dataset. He has put forward human activity recognition pipelines capable of adaptation, of exploiting opportunistic sensing, and more recently capable of lifelong learning for open-ended activity recognition. He received his Master's degree (2001) and PhD (2005) from the Ecole Polytechnique Federale de Lausanne, Switzerland.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.